Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 604 – 611

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Analysis of Stylometric Variables in Long and Short Texts

Fernanda López-Escobedo, Carlos-Francisco Méndez-Cruz, Gerardo Sierra*, Julián Solórzano-Soto

Instituto de Ingeniería-Universidad Nacional Autónoma de México,Circuito Escolar s/n Ciudad Universitaria, México D.F 04510, México

Abstract

This paper presents some experiments in the task of authorship attribution. We achieve this task by a stylometric analysis of some stylistic markers tested in two Spanish corpora. The first corpus is composed of long texts written by professional authors, while the second corpus is formed by short texts written by students. In both corpora, different text genres are included. Thus, the objective of this study is to analyze several stylometric variables to test its capacity as markers for authorship attribution when the corpora vary in size and text genre. We represent the texts as high dimensional vectors and we visualize the similarities between them using multidimensional scaling. We conclude that the length of texts is a factor that affects the discriminatory capacity of the stylometric variables. We also found that there are certain variables that are better than others to identify specific authors and specific text genres.

Keywords: authorship attribution; stylometry; stylometric variables; multidimensional scaling

1. Introduction

One of the main approaches to authorship attribution is corpus-based stylometric analysis. Corpora used in this task are commonly formed by a set of documents of several authors. As in every corpus-based approach, corpus design represents a key aspect in order to produce feasible results of analysis. In the case of authorship attribution, corpus features (text genre, dialect, era, register) have impact in the precision of task of attribution. What is more, it has been proposed that "it is best to compile author-based corpora that represent the narrowest variety of language possible" (Grieve, 2007).

* Corresponding author. Tel.: +52-555-623-3600; fax: +52-555-623-3507.
E-mail address: gsierram@iingen.unam.mx

doi:10.1016/j.sbspro.2013.10.688

Because of the growing use of social media and portable devices, which impose relative limits on message size, the interest of studying how the corpus size impacts on authorship attributions has increased in recent years. For example, attributing tweets to writers represents a difficult challenge due to 140-character limit. Consequently, our interest is in experimenting with different corpus sizes.

With regard to the "correct" size of a corpus, it is hard to know how large or how short the corpus should be for authorship attribution. In contrast, other kind of corpus-based studies, such as lexicographic studies, can approximate the size of the corpus based on, for example, type/token ratio. So, in this work we use two corpora, one with long texts (10000 words) and one with short texts (500 words) to represent different corpus sizes.

First of all, we present some main concepts related with authorship attribution and stylometry. Then, we expose the objective of our study. Later, the methodology is mentioned followed by the experiments and results. Finally, we provide some conclusions and future work.

1.1. Authorship Attribution

Determining the authorship of an anonymous text, or to solve a controversial authorship of a set of documents has been a long-standing human concern. Indeed, the interest in authorship attribution comes not only from literary areas, but also from legal ones. According to Juola (2006) we can roughly define authorship attribution as the task of inferring the author of a document.

According to the research done by Koppel, Schler & Argamon (2009) there are at least three types of authorship attribution problems:

- There are many candidates and we have to attribute the text to one of them.
- There is one suspect and we must determine if the suspect is the author of the text or not.
- There are no suspects. The task is to provide as much psychological or demographic information about the author as possible.

Different methods have been used throughout history to approach the authorship attribution problem. Mendenhall (1887) and Mascol (1888) are two of the pioneers in the field. This early work proposed that the writing of each author could be characterized by a curve expressing word length distribution. The case of the Federalist Papers done by Mosteller and Wallace (1964) is often cited as the beginning of modern work in authorship attribution which introduced a multivariate analysis approach. They employed Bayesian classification using the frequencies of certain function words as features. Among other commonly used multivariate statistics methods for authorship attribution nowadays we can mention Linear Discriminant Analysis (Bayen, van Halteren, Neijt, & Tweedie, 2002), Principal Component Analysis (Jamak, Savatić, & Can, 2012) and Karhunen-Loeve transforms (Abbasi & Chen, 2008).

Also, data visualization techniques have been used to try to create textual "fingerprints" that represent the style of an author. Keim & Oelke (2007) create a visual representation of a document by assigning colors to each section of the text according to various measures. Abbasi & Chen (2006) invented a system called "Writeprints" which creates a distinctive 3D pattern for each author based on a sample of their texts.

1.2. Stylometry

As for the technique to resolve the task of authorship attribution, we use stylometry. It means that for our purpose, authorship attribution and stylometry are not equivalents, although some authors have proposed the equality of these terms (Juola, Sofko, & Brennan, 2006).

We understand stylometry as those techniques that allow measure the style of an author by the identification of its features of style (stylemas). Those stylemas, also called style (stylistic) markers (Stamatatos, 2009), are obtained from textual measurements normally calculated by statistical methods.

According to Madigan, Genkin, Lewis, Argamon, Fradkin & Ye (2005), the most popular style markers are the so-called function words (such as 'a', 'the', 'of'), because they are considered to be topic independent. Other stylometric features that are commonly used include various measures based on vocabulary richness (which aren't very reliable due to their dependence on the length of the text), word class frequencies, word collocations, grammatical errors and word, sentence and paragraph lengths.

Nowadays, stylometry has also incorporated Natural Language Processing (NLP) methods to explore different style markers based on syntactic analysis. In this paper, we combine both methods, particularly a Part of Speech (POS) technique from NLP.

In general, after the style markers of a document are obtained, they are compared to style markers of different documents of several potential authors. Authorship attribution is reached when a best match is established. One of the main problems of this approach is the definition of a set of markers which delivers significant results. Regarding that problem, there are some studies about numerous style markers and its relevance in the task of attribution. One of them was made by Grieve (2007). For Spanish another was made by Blasco & Ruiz (2009).

2. Objective

As outlined earlier, the size of a text could represent a factor affecting the discriminatory capacity of stylistic markers. Therefore, in this paper two corpora of different sizes are studied. In addition, different text genres are tested, considering that this characteristic could affect the stylometric variables as markers for authorship attribution.

Thus, the objective of this work is to analyze seven stylometric variables in two different corpora (different in size and text genre) in order to test its capacity as markers for authorship attribution.

3. Methodology

3.1. Stylometric variables

A lot of stylometric variables have been used by different researchers in several studies. On average, 20 variables are used in each experiment (Abbasi & Chen, 2008, De Vel, Anderson, Corney, & Mohay, 2001, Grieve, 2006, Koppel & Schler, 2003). Normally, their inclusion depends on the domain of the application. For example, methods for authorship attribution for e-mails and other short online texts take into account structural style markers such as the presence of greetings, file attachments, certain HTML tags, etcetera, or idiosyncratic style markers such as spelling mistakes (Abassi & Chen, 2008, Koppel & Schler, 2003). Being literary writing the domain of our experiment, we opted to include the more general style markers. The stylometric variables we chose are the following:

- Punctuation (individual occurrence of 20 punctuation marks)
- Function words n-grams (unigrams, bigrams and trigrams, with and without gaps)
- Content words n-grams (unigrams, bigrams and trigrams, with and without gaps)
- POS tags n-grams (unigrams, bigrams and trigrams, with and without gaps)
- Word length frequency distribution
- Type token ratio
- Hapax legomena count

3.2. Text Representation

For each stylometric variable, we generate a frequency vector where every dimension corresponds to a different feature. In all experiments, we choose a different combination of variables and then we represent each document as the concatenation of the frequency vectors corresponding to those variables.

3.3. Multidimensional Scaling

Classical multidimensional scaling (CMD) is a statistical technique for data visualization. It has been previously used for authorship attribution by other researchers such as Merriam (2003). For a set of N objects, it takes as input a NxN matrix in which every element $d_{i,j}$ represents the dissimilarity between the i-th and j-th object. Then it assigns a

location in a *p*-dimensional space (for a previously chosen *p*) to each object, so that the distance between them in this space, is equivalent to their distances in the dissimilarity matrix.

In this case, objects are documents, and the dissimilarity between each of them is represented by the Euclidean distance between their feature vectors. Using this technique, 2-dimensional scatter plots were obtained, with each point representing each text. The distance among points denotes its similarity in a relatively easy way to visualize.

3.4. Corpus design

Our Spanish corpus consists of 27 long texts written by professional authors (10,000 word tokens on average), and of 15 short texts made by students (500 word tokens on average). The long texts consist of a variety of textual genres: journalism, novel, essay, play, and short story. Nine texts per each of three professional authors are chosen and for all of them we select three different genres (Table 1). Regarding the short texts, there are five authors, and for each one there are three texts: a piece of fiction, an emotive anecdote, and an argumentative essay (Table 2).

This variety was considered in an attempt to investigate whether certain combinations of stylometric variables can represent the style of an author even if their texts are of different textual genres.

Table 1. Corpus design (long texts)

Author	Work	Genre	Alias
Jorge Ibargüengoitia	Dos crímenes	Novel	JI1
Jorge Ibargüengoitia	Estas ruinas que ves	Novel	JI2
Jorge Ibargüengoitia	Instrucciones para vivir en México	Journalism	JI3
Jorge Ibargüengoitia	La Casa de usted y otros viajes	Journalism	JI4
Jorge Ibargüengoitia	La ley de Herodes	Short story	JI5
Jorge Ibargüengoitia	Las muertas	Novel	JI6
Jorge Ibargüengoitia	Los pasos de López	Novel	JI7
Jorge Ibargüengoitia	Los relámpagos de agosto	Novel	JI8
Jorge Ibargüengoitia	Maten al león	Novel	JI9
Jorge Luis Borges	Borges en Sur	Journalism	JLB1
Jorge Luis Borges	El Aleph	Short story	JLB2
Jorge Luis Borges	El Libro de Arena	Short story	JLB3
Jorge Luis Borges	Ficciones	Short story	JLB4
Jorge Luis Borges	Historia de la Eternidad	Essay	JLB5
Jorge Luis Borges	Inquisiciones	Essay	JLB6
Jorge Luis Borges	Otras Inquisiciones	Essay	JLB7
Jorge Luis Borges	Revista Multicolor	Journalism	JLB8
Jorge Luis Borges	Textos Publicados En El Hogar	Journalism	JLB9
Mario Vargas Llosa	El loco de los balcones	Play	MVL1
Mario Vargas Llosa	El pez en el agua	Essay	MVL2
Mario Vargas Llosa	El sueño del celta	Novel	MVL3
Mario Vargas Llosa	Kathie y el hipopótamo	Play	MVL4
Mario Vargas Llosa	La ciudad y los perros	Novel	MVL5
Mario Vargas Llosa	La civilización del espectáculo	Essay	MVL6
Mario Vargas Llosa	La orgía perpetua	Essay	MVL7
Mario Vargas Llosa	La señorita de Tacna	Play	MVL8
Mario Vargas Llosa	Travesuras de una niña mala	Novel	MVL9

Table 2. Corpus design (short texts)

Author	Genre	Alias
CP001	Fiction	CP001-1
CP001	Emotive	CP001-2
CP001	Argumentative	CP001-3
CP002	Fiction	CP002-1
CP002	Emotive	CP002-2
CP002	Argumentative	CP002-3
CP003	Fiction	CP003-1
CP003	Emotive	CP003-2
CP003	Argumentative	CP003-3
CP004	Fiction	CP004-1
CP004	Emotive	CP004-2
CP004	Argumentative	CP004-3
CP005	Fiction	CP005-1
CP005	Emotive	CP005-2
CP005	Argumentative	CP005-3

4. Experiments and results

Relative frequencies for all style markers for each text were generated by means of a tool programmed in the Python language. Then another tool was created which took as input the combination of variables that we wanted to analyze, and generated a data matrix made up from the frequencies of the selected features. This data matrix was given to an R script (R Development Core Team, 2011) which calculated the dissimilarity matrix using the Euclidean distance between the vectors. Then, it performed a CMD analysis over the matrix, and presented the resulting plot.

Long and short texts were analyzed separately by different combinations of variables. We can see that in the first plot (Fig 1.a) the three authors were more clearly separated than in the second one. Also, in neither plot the texts of Mario Vargas Llosa were grouped together. However, it is worth noting that in both cases the texts MVL1, MVL4, and MVL8 (shown inside squares), which are all three of his plays, appear close to each other (more closely in the second plot, where punctuation marks are taken into account) and far from the rest.

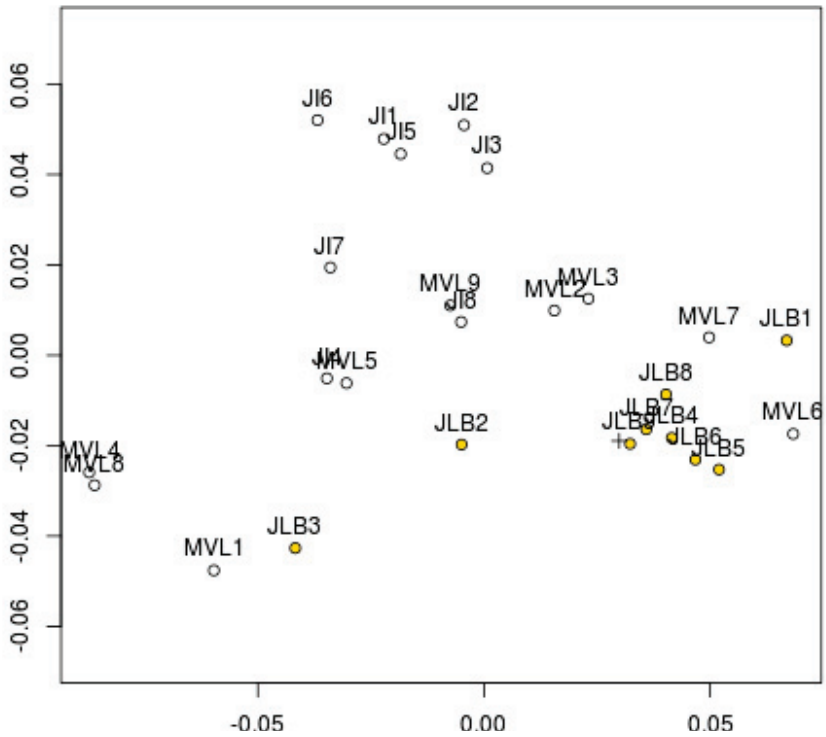
On the other hand, in the case of the short texts (Fig 2), it was more difficult to separate the authors using combinations of our selected variables.

5. Conclusions and future work

First and foremost we could see that experiments for short tests were less satisfactory than for long texts, since the variables couldn't clearly form clusters for the authors. This leads us to the conclusion that the length of the texts is a very important factor to consider when choosing the stylometric variables to be used.

It was also observed that while some combination of variables may be very good at grouping the texts of a certain author (the case of Ibargüengoitia and Borges), it may be at the same time very bad at grouping the texts of a different author (Mario Vargas Llosa). It is worth noting also that the plays of Mario Vargas Llosa tended to be farther from the rest of his texts. All of this means that there are genres and variables which will cause the texts of one same author to appear considerably away from each other and so they must be found prior to attempting authorship attribution. In other words, the best combination of variables depends on each specific author.

a)



b)

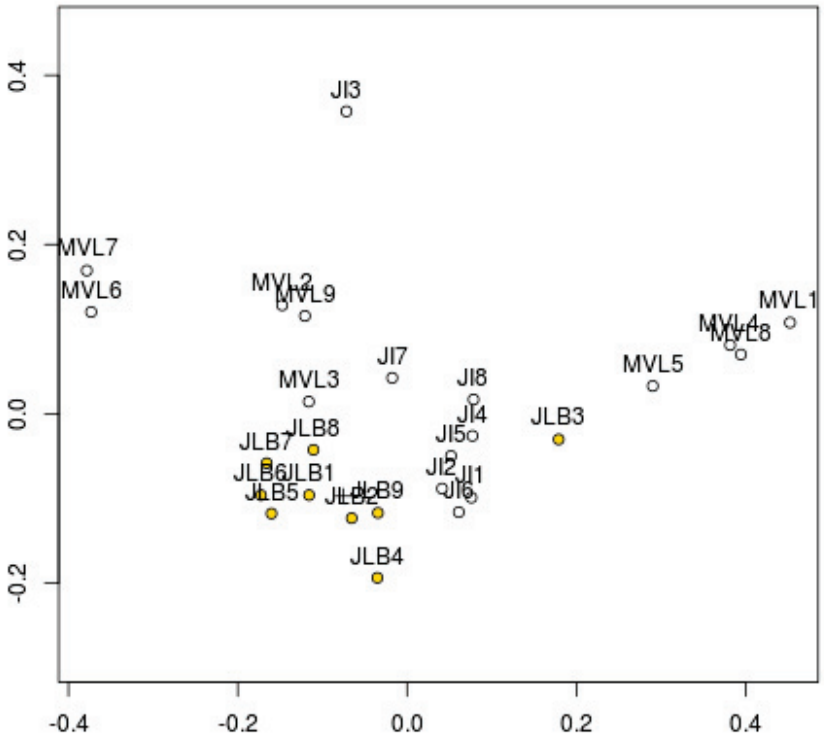
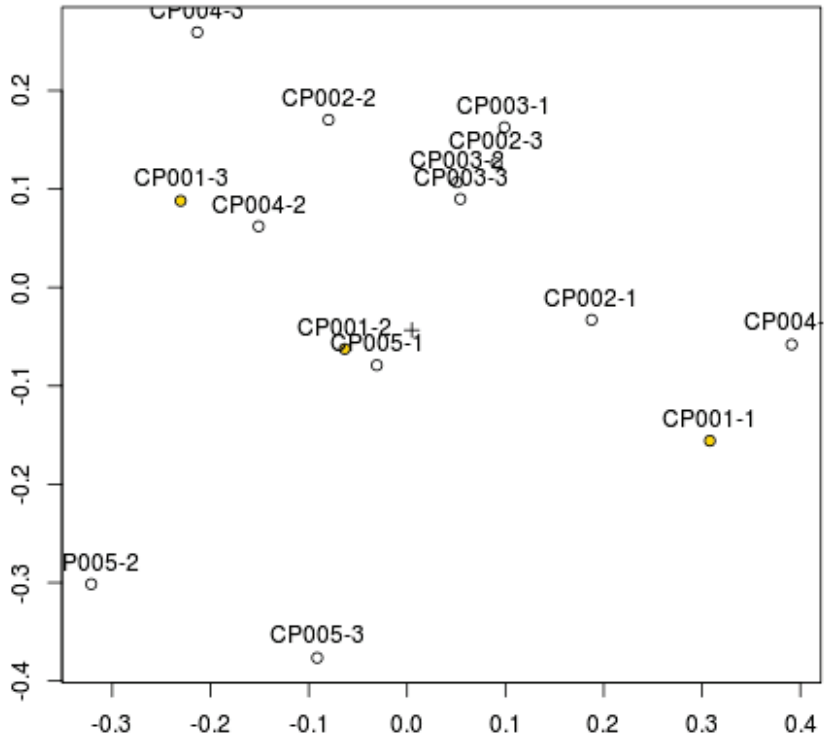


Fig. 1. (a) Trigrams of POS tags and trigrams of function words, long texts; (b) Punctuation marks, type token ratio, word length distribution, POS tag trigrams, function words trigrams, long texts.

a)



b)

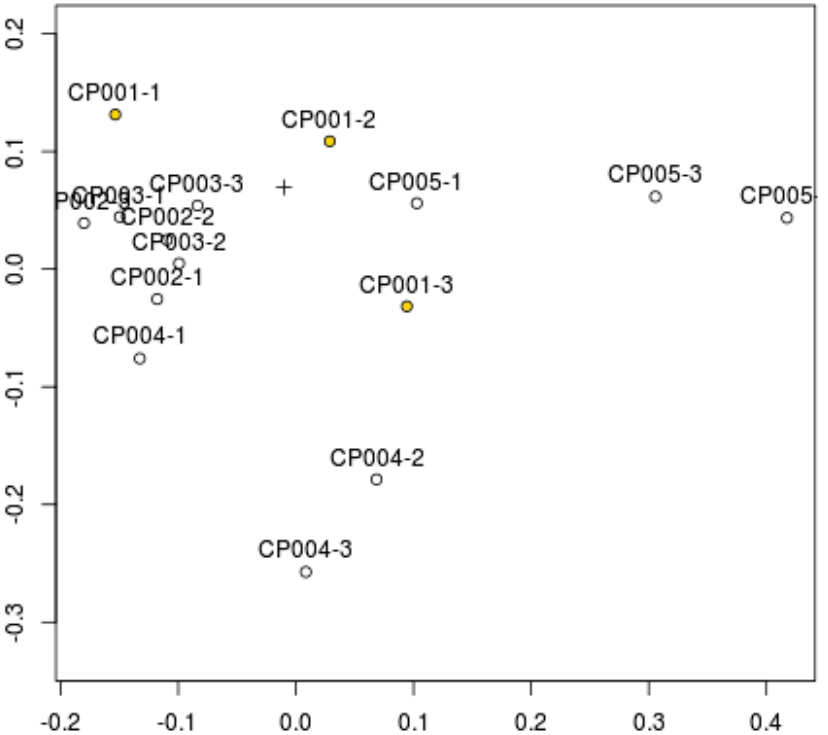


Fig. 2. (a) Punctuation marks, type token ratio, word length distribution, POS tag trigrams, function words trigrams, short texts; (b) Punctuation marks and content words unigrams, short texts

Some questions remain open such as how to find the best combination of variables. Currently we are developing an algorithm to solve this as an optimization problem, i.e. systematically trying many combinations and, according to a well defined metric, selecting the one with the best score.

The number of combinations that can be tried is very large. In this case we used 7 stylometric variables but counting the unigrams, bigrams and trigrams with and without gaps as 5 variants of each variable, we actually have 19 variables to combine. If we were to try every possible combination, we would have to consider all combinations of two variables, all combinations of three variables, and so on. The total number of combinations can be computed as 2^n. This is because we can consider the combination of variables as a vector of 19 bits. Each bit corresponds to a different variable, and thus each one of them can be either "on" or "off". So, for n=19 the total number of combinations is 524,288. There are, of course, some combinations that can be discarded immediately because certain variables are mutually exclusive (such as n-grams with gaps and n-grams without gaps), and we could also restrict the combinations to only certain number of "on" bits. It is still a large number so an algorithm must be devised to efficiently try most of them.

Finally, it can be debated whether the two corpus are comparable the way we intended to. It has been suggested that the short texts should be texts written by the same authors than the long texts. This is feasible since many professional authors write blogs or microblogs, besides normal literary texts such as novels. These could be used to construct both corpora instead of using different authors for the long and short texts.

Acknowledgements

We would like to acknowledge the sponsorship of the project PAPIIT-UNAM IN400312 "Análisis estilométrico para la detección de similitud textual", as well as CONACYT CB2012/178248 "Detección y medición automática de similitud textual"

References

Abbasi, A., & Chen, H. (2006). Visualizing authorship for identification. *Intelligence and Security Informatics* (pp. 60-71). Springer Berlin Heidelberg

Abbasi, A. & Chen H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS),* 26(2), Article 7.

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). *An experiment in authorship attribution (6th JADT)*, 29-37.

Blasco J., & Ruiz C. (2009). Evaluación y cuantificación de algunas técnicas de atribución de autoría en textos españoles. *Castilla: Estudios de Literatura*, 27-47.

De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4), 55-64.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques, *Literary and Linguistic Computing*, 22 (3), 251-70.

Jamak, A., Savatić, A., & Can, M. (2012). Principal component analysis for authorship attribution. *Business Systems Research*, 3(2), 49-56.

Juola, P. (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1 (3), 238-239.

Juola, P., Sofko, J., & Brennan, P. (2006). A prototype for authorship attribution studies. Literary and Linguistic Computing, 21(2), 169-178.

Keim, D. A., & Oelke, D. (2007). Literature fingerprinting: A new method for visual literary analysis. *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium* on (pp. 115-122). IEEE.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 69, 72-80.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), 9-26.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. *Proc. of the Meeting of the Classification Society of North America*.

Mascol, C. (1888a), Curves of pauline and pseudo-pauline style i. *Unitarian Review, 30:452-460*,1888.

Mascol, C. (1888b), Curves of pauline and pseudo-pauline style ii. *Unitarian Review, 30:539-546*,1888.

Mendenhall, T. C. (1887), The characteristic curves of composition. *Science 9,* pp. 237-249.

Merriam, T. (2003). An application of authorship attribution by intertextual distance in English. *Corpus*, (2).

Mosteller, F., Wallace, D. L. (1964), Inference and Disputed Authorship: *The Federalist. Reading, Mass. Addison Wesley*

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), 538-556.